Maintaining Data Quality in Data Warehouse

Rahul Gupta

Master of Business Administration (Computer Information Systems), University of Rochester, NY

Abstract- Data warehouse forms an integrated environment where data from disparate systems is bought together and presented in a consistent matter. Based on the data present in the warehouse business users want to make key decisions. In order to support these decisions, quality of data in warehouse must be reliable. If there are issues in data loaded in warehouse, business users lose trust and the information from the warehouse becomes unreliable. One of the essential components of data warehouse is the (Extract, Transform and Load) ETL process. This process is used in most organizations to load data into the warehouse. An ETL is a complex process in itself and one which is most time consuming during the building of the data warehouse. It allows the data in the warehouse to be refreshed on the periodic basis (daily, 'maintained and there is no data loss every time data is loaded into the warehouse using ETL.

An audit balance and control (ABC) framework is for this purpose. Using this framework quality of data in warehouse can be maintained.

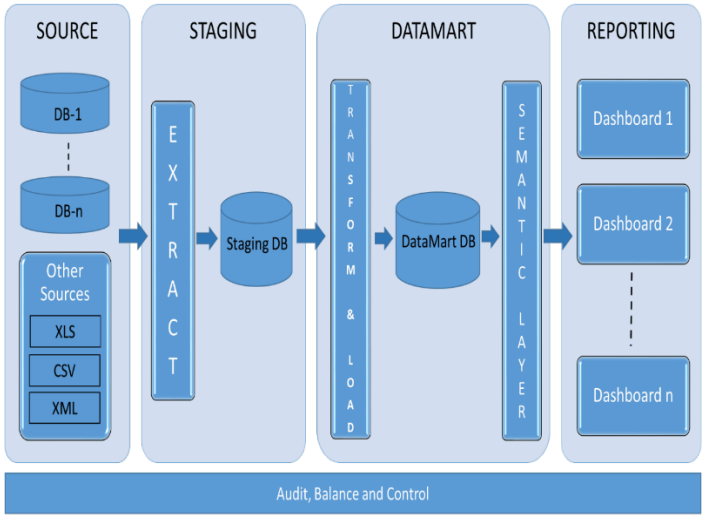
Each of the three layers (Audit, Balance and Control) serve a specific purpose in the data warehouse. Using Audit the schedule of various ETL jobs that run on daily basis, job start time and job end time, number of records loaded, number of records that errored outcan be traced. Balance is for comparing the source data with the data loaded into the staging and into the rules to be applied, and have the functionality to report on exceptions.Control is for scheduling the ETL jobs and re-starting the ETL jobs in case of failures.

Index Terms- Audit Balance and Control (ABC), ETL, Data Quality, Data Warehouse

I. Introduction

In this framework, Audit is for tracking many of the processes used for ETL maintenance and operations. Not only the errors can be reported but also the quality of data loaded into the warehouse can be checked.

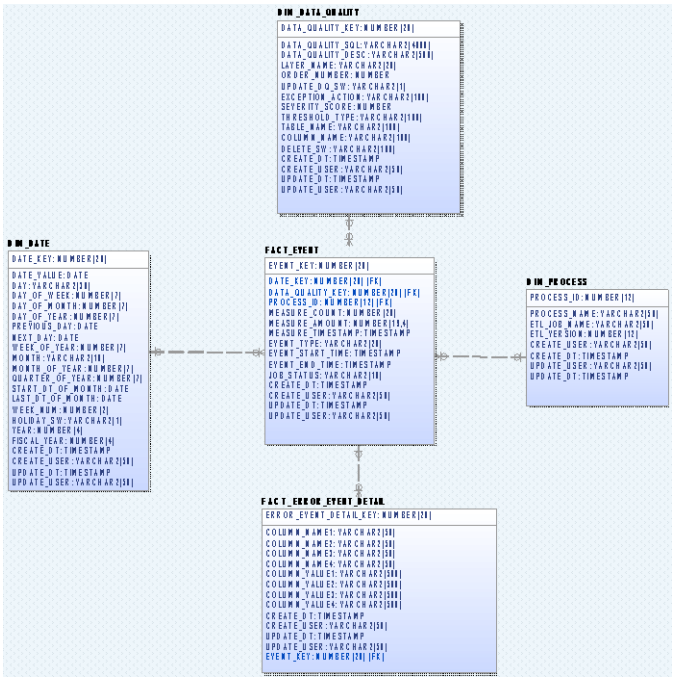
A.



As illustrated in diagram above, Audit Balance and Control framework can be used to audit balance and control the data movement between different layers such as source to staging and staging to DataMart.

1. The following are the main technical features which can be implemented by ABC subject area:-
2. Audit of all ETL batch jobs that run i.e. one event generated for each job start and job end with an indication if the job completed successfully, completed with warnings, or failed.
3. A predefined set of sql's are run at specific points in the ETL process. For example sql's are run sql' s at source, staging, and DataMart layers and the count of records compared.
4. Ability to control the ETL batches that run and to start or stop a process based on the output of the sql from step 2.
5. Perform row level exception handling i.e. flagging the problem records or records with data quality issues in the DataMart so they could be excluded in reporting.

Below is the data model for implementation of Audit Balance and Control (ABC) DataMart:-



DIM_DATE:-This is standard date dimension used to track date when the event occurred. For example, the date when the audit event was recorded.

DIM_PROCESS:-This dimension is used to store the information for ETL batch jobs. Using this dimension, name of each ETL job, its description and version can be tracked.

DIM_DATA_QUALITY:- This table is used for data quality checks which can be implemented using sql statements. This stores the sql scripts which are run against the source, staging or data warehouse layer to keep a count of the records or to compare the metrics in each layer. It can also be used for row level check to flag the problem records in each layer.

FACT_EVENT:-This table records the date when the event occurred along with the record count in case of the control event. In case of audit event, it stores the date when event occurred along with count of the records which errored out.

FACT_ERROR_EVENT_DETAIL:-This table stores the row level details of the records which errored out. The records are stored in key value pair where key represent the column name and value represents the actual value of the column.

Implementation of Audit:-

In audit, we basically store the start and end times of each job along with how many records were processed in each run. We also store the records that errored out during the loads.

To support the requirements of the audit a separate DIM_PROCESS table which has names of all the jobs is created. Each time a job is run, the start and end time of the job along with number of records processed is stored in FACT_ABC_EVENT. This way we can keep track of what all jobs ran successfully and what failed.

When an ETL job is run the successful records are loaded into respective target tables while the rejected recordsare stored in FACT_ERROR_EVENT_DETAIL table. This table holds the rejected records in key value pair form along with error message and error timestamp. The key values pair provides flexibility to include any tables without doing any structural changes in the error table.

Implementation of Balance: -

In Balance, the ETL process is run so as to compare the results between different layers such as Source, Staging and DataMart. To support this, DIM_DATA_QUALITY will have list of sql scripts which will be stored and run at pre-defined intervals compare data in different layers. There can be two types of check done using this:-

1) The check to compare counts in different layers such as source, staging and DataMart.
2) The check to compare aggregates such as Amounts etc.

Implementation of Control:-

In control, we maintain the schedule of ETL jobs as well as the re-startability of ETL jobs in case of failures. Using the D_DQ_SCRIPT tables, we can also do row level checks and control the quality of the data loaded into the DataMart.

ETL job failures will be identified by querying the FACT_ABC_EVENT table based on the status attribute value as 'Failure'. Manual intervention will be required to identify the issue, fix it, and rerun the job. Issues can be database, source file structure incorrect, invalid data etc. and based on this analysis appropriate action needs to be taken and rerun the job in case of fixing the issue. In case job being success with rejected records, rejected records captured in the error table must be analyzed and correct record need to be sent in the next run to reprocess.

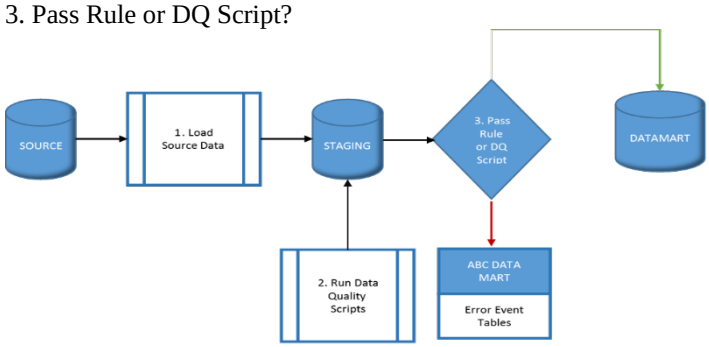
For doing row level checks and controlling the quality of data loaded into DataMart following process is followed:-

First the data is loaded from source into staging area. The data quality SQL Scripts is run on the Staging Table. The scripts should dynamically read from DIM_DQ_SCRIPT and execute the DQ process. It will update the DQ Flag, DQ_Severity and DQ_ActionCode appropriately on the Stagingtables and determine DQ Flag = P/F (Pass/Fail).

If the DQ Flag = 'F' OR DQ_ActionCode = "Not Load in DM" then records get loaded into the FACT_ERROR_EVENT_DETAIL. If the DQ_FLAG = 'P' and DQ_ActionCode = "Load in DM" then it gets loaded in the DataMart.

The below diagram demonstrates the above data quality mechanism:

3. Pass Rule or DQ Script?
3. Pass Rule or DQ Script?
3. Pass Rule or DQ Script?



II. RESULT FINDINGS

The table below details the actions being taken from the numbered steps in the diagram above.

1. Source data is loaded into the Staging Layer via ETL process
2. Data Quality scripts held within the ABC Data Mart's are run against the staging layer
3. Records that meet data quality requirements PASS, if not, the records FAIL
4. Clean Record Set will be loaded into the Data Marts
5. Failed Records will be loaded into the ABC Data Mart's Error Event Tables

III. CONCLUSION

In conclusion the reconciliation of data from source to warehouse is critical to overall success of the business intelligence program. Only if the data is reconciled on day to day basis can the business have faith in the report produced from the data warehouse. Using the Audit Balance and Control the quality of data in data warehouse can be tracked and maintained. To thisextent, ABC framework is helpful so as to have trust of business on the data.

AUTHORS

First Author –Rahul Gupta, Master of Business Administration (MBA), Simon School, University of Rochester,mitarahul79@gmail.com.

Correspondence Author –Rahul Gupta, Master of Business Administration (MBA), Simon School, University of Rochester,mitarahul79@gmail.com,617-416-9502.